KLASIFIKÁCIA ZAŠUMENÝCH DÁT

ANALÝZA DYCHU

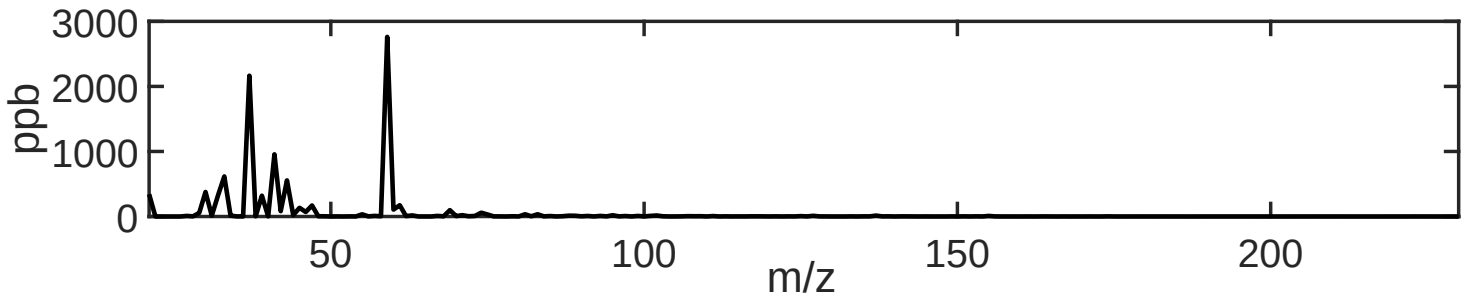
Katarína BARTOŠOVÁ

Analýza dychu ako diagnostický nástroj

- PTR-MS, hmotnostná spektrometria s protónovou prenosovou reakciou
  - meranie nízkych koncentrácií prchavých organických zložiek VOCs vydychovaného plynu
- Analýza dychu
  - neinvazívna diagnostická metóda
  - potenciál včasne diagnostikovať rakovinu pľúc alebo pažeráka
  - včasným diagnostikovaním sa môže poskytnúť skorá liečba, vďaka čomu môžu byť zachránené tisíce životov ročne

Profil vydychovaného vzduchu

- koncentrácie - častica na miliardu, 1 ppb $= 1\ \mu$g/kg, 10^{-9}
- molekuli s pomerom molekulovej hmotnosti k náboju od m/z 21 po m/z 230 $\Rightarrow$ 210 zložiek
- m/z ako VOC s najsilnejším zastúpením, napr. m/z 42 ako acetonitril



Analyzované dáta

- profily pre 219 "zdravých"subjektov
  - viacero vzoriek s časovým odstupom
  - každá vzorka meraná minimálne tri krát

- pre každý subjekt berieme len jeden reprezentatívny vektor vypočítaný ako medián z mediánov opakovaných meraní $\Rightarrow$
  dáta v sebe zahŕňajú variabilitu opakovaných meraní

ZAŠUMENÉ DÁTA

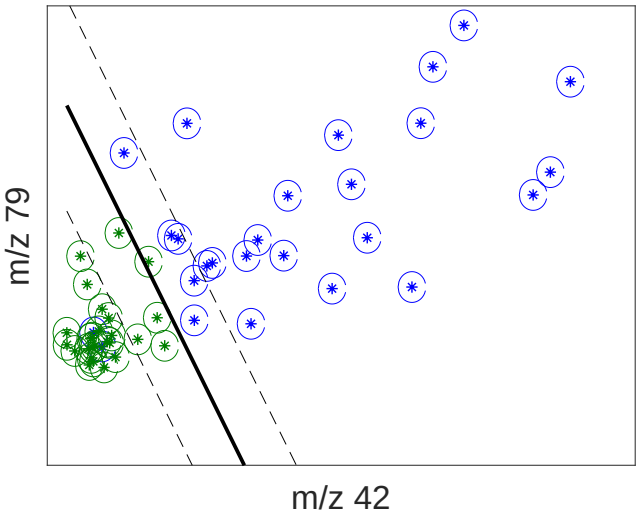
Klasifikácia zašumených dát

- analýza dychu ako diagnostický nástroj
  - neodmysliteľný vzťah, rakovina pľúc a fajčenie
- klasifikácia "zdravých"dobrovoľníkov na základe fajčiarskeho návyku
  - skupina fajčiarov, $n^{(1)} = 44$
  - skupina nefajčiarov, $n^{(2)} = 173$

Klasifikácia zašumených dát

- robustná formulácia metódy oporných bodov RSVM
- elipsoidálny model šumu,
  tzn. že skutočná hodnota, nie vždy aj nameraná hodnota je nejaký bod v špecifikovanom elipsoide
- nie je nutný predpoklad typu rozdelenia pozorovaných dát, predpokladáme len konečnosť momentov druhého rádu
- ROBUST 2008 a ROBUST 2010

Klasifikácia zašumených dát



RSVM

- hľadáme rozh. funkciu

$$g(\mathbf{x}) = \langle \mathbf{w}, \mathbf{x} \rangle + b$$

na základe špecifikovaných elipsoidov $\mathbf{x} \in B(\mathbf{x}_i, \Sigma_i, \gamma)$

- $\mathbf{x}_i, \Sigma_i$ sú stred a matica tvaru šumu pre i-ty elipsoid
- γ je nastaviteľný parameter hladiny zašumenia, $\gamma \geq 0$, $\gamma = 0$ bez šumu
- optimalizačná úloha sa rieši ako úloha kónického programovania druhého rádu

pre rastúci počet vstupných premenných

- profil vydychovaného vzduchu pre klasifikované subjekty charakterizovaný zvyšujúcim sa počtom VOCs
- 2 VOCs, 3 VOCs, 4 VOCs, ..., 210 VOCs
  - od najmenej významých po štatisticky významnejšie
  - od najvýznamnejších VOCs po menej významné

Štatisticky významné VOCs

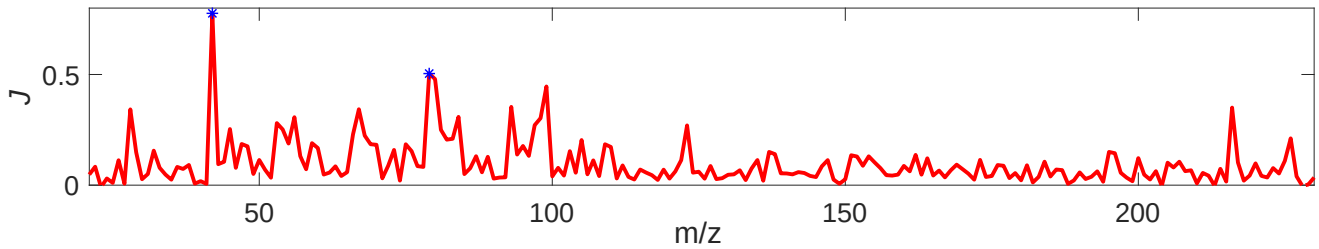
- efektívnosť zatriedenia subjektov podľa sledovaného znaku - Youdenov index

$$J = \max_t \{Se(t) + Sp(t) - 1\}$$

- pre $\forall t$ z X
- $Se(t) = 1 - F^{(1)}(t)$
- $Sp(t) = F^{(2)}(t)$

- pretože empirické distribučné funkcie niesú spojité, odhad J sa správa veľmi nepredvídateľne, obzvlášť keď $n^{(1)} \neq n^{(2)}$ $\Rightarrow$ sme použili vyhladené odhady pomocou funkcie Gaussovských jadier

Štatisticky významné VOCs



- najefektívnejšie zložky pre fajčiarsky návyk m/z 42 - acetonitril a m/z 79 - benzén

Simulačná štúdia

- klasifikačné pravidlo sme natrénovali na rovnako zastúpených skupinách,
  26 fajčiarov a 26 nefajčiarov; 60% z nemej početnej skupiny
- na zvyšných subjektoch sme klasifikačné pravidlo vyhodnotili
  18 fajčiarov a 147 nefačiarov
- výsledky simulačnej štúdie sme zobrali ako priemer pre 100-krát náhodne rozdelené jednotlivé skupiny subjektov na trénovaciu a testovaciu skupinu

Simulačná štúdia

- $\hat{Se} = P(\hat{y} = +1|y = +1) = \frac{\#\{i, y_i = \hat{y}_i | y_i = +1\}}{\#\{i, y_i = +1\}}$

  schopnosť klasifikátora rozpoznať prítomnosť sledovaného znaku v pozitívnej triede

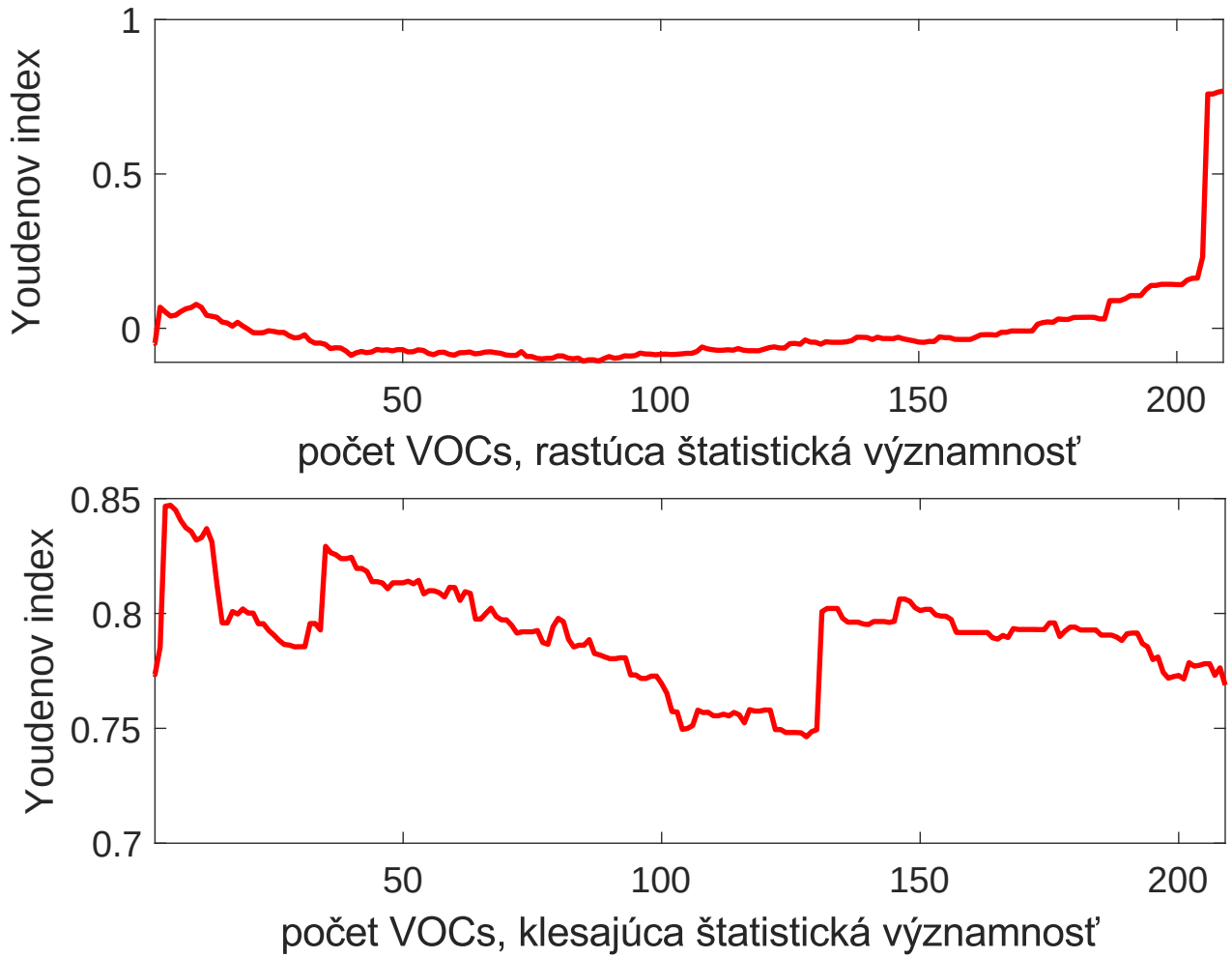
- $\hat{Sp} = P(\hat{y} = -1|y = -1) = \frac{\#\{i, y_i = \hat{y}_i | y_i = -1\}}{\#\{i, y_i = -1\}}$

  nakoľko klasifikátor správne zatriedi subjekty, kt. vlastnosť skutočne nenesú

- $\hat{J} = \hat{Se} + \hat{Sp} - 1$

  efektívnosť klasifikácie

Simulačná štúdia



Ďakujem za pozornosť!